AMERICAN SOCIETY FOR MICROBIOLOGY | Microbiology Resource Announcements

GENOME SEQUENCES

Draft Genome Sequences of Five Fungal Strains Isolated from Kefir

Simonas Marcišauskas,[a] Yongkyu Kim,[b]* Sonja Blasche,[b]* Kiran R. Patil,[b]* Boyang Ji,[a,c] Jens Nielsen[a,c]

[a]Department of Biology and Biological Engineering, Chalmers University of Technology, Gothenburg, Sweden
[b]European Molecular Biology Laboratory, Heidelberg, Germany
[c]BioInnovation Institute, Copenhagen, Denmark

ABSTRACT We present the annotated draft genome sequences of five fungal strains isolated from kefir grains. These isolates included three ascomycetous (*Candida californica*, *Kazachstania exigua*, and *Kazachstania unispora*) and one basidiomycetous (*Rhodotorula mucilaginosa*) species. The results revealed a detailed overview of the metabolic features of kefir fungi that will be potentially useful in biotechnological applications.

Kefir is fermented milk traditionally produced by a specific symbiotic culture of bacteria and fungi. Also known as kefir grains, this culture usually consists of 40 to 50 different species, including lactic acid bacteria, acetic acid bacteria, and yeasts (1). The ascomycetous yeast *Kluyveromyces marxianus* was previously identified in kefir grains (2), but little is known about other cooccurring fungi. Here, we report the annotated whole-genome sequences of the ascomycetous yeasts *Candida californica*, *Kazachstania exigua*, and *Kazachstania unispora* and the basidiomycetous fungus *Rhodotorula mucilaginosa*, isolated from kefir grains collected from private sources. These kefir grain cultures were collected in Germany (Ger04, *C. californica* and *K. unispora*; Ger06/OG2, *K. exigua*) and South Korea (Kefir Korea, *R. mucilaginosa*). *C. californica* SB-48 (referring internal stock identifier) was isolated from ground kefir grains and plated in serial dilutions onto yeast extract-peptone-dextrose-adenine (YPDA) medium. *C. californica* SB-116 was isolated and plated in serial dilutions onto Sabouraud dextrose (SD) medium. *K. exigua* SB-178 was isolated and plated in serial dilutions onto M17 medium supplemented with glucose. *K. unispora* SB-162 was isolated and plated in serial dilutions on de Man, Rogosa, and Sharpe (MRS) agar-milk agar (1/1 mix of MRS agar and 3.5% ultrahigh-temperature processing [UHT] milk). *R. mucilaginosa* SB-353 was isolated and plated in serial dilutions onto tomato juice agar (TJA). All isolates were grown in their corresponding medium for up to 5 days at 30°C. Isolates were identified by internal transcribed spaced (ITS) DNA amplification PCR using the primers S-D-Bact-0515-a-S-16 (GTGCCAGCMGCNGCGG) and S-*-Univ-1392-a-A-15 (ACGGGCGGTGTGTRC) (3) and subsequent Sanger sequencing of the amplified region. ITS sequences were taxonomically assigned using an open-reference method. The kefir-isolated yeast was used as the reference, and subsequent naive Bayesian classification was performed using UNITE (4). Strains were deposited and are available in the Leibniz Institute DSMZ collection of microorganisms under the same strain names.

The genomic DNA extraction was performed using a two-step approach combining enzymatic digestion with lysozyme, followed by bead beating with 0.3-g glass beads. The supernatant was then digested with proteinase K and applied to phenol-chloroform extraction and DNA precipitation, as described in references 5 and 6. DNA was then prepared for sequencing using a Nextera DNA library preparation kit (Illumina) and sequenced on an Illumina HiSeq 2000 instrument to get 100-bp paired-end reads with the insert size ranging between 250 bp and 300 bp. The quality of reads was checked with FastQC v0.11.9 (7), while Trimmomatic v0.36 (8) was used to adapter and

Citation Marcišauskas S, Kim Y, Blasche S, Patil KR, Ji B, Nielsen J. 2021. Draft genome sequences of five fungal strains isolated from kefir. Microbiol Resour Announc 10:e00195-21. https://doi.org/10.1128/MRA.00195-21.

Editor Antonis Rokas, Vanderbilt University

Copyright © 2021 Marcišauskas et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Address correspondence to Jens Nielsen, nielsenj@chalmers.se.

* Present address: Yongkyu Kim, Brain Research Institute, Korea Institute of Science and Technology, Seoul, South Korea; Sonja Blasche, Medical Research Council (MRC) Toxicology Unit, University of Cambridge, Cambridge, United Kingdom; Kiran R. Patil, Medical Research Council (MRC) Toxicology Unit, University of Cambridge, Cambridge, United Kingdom.

Received 23 February 2021
Accepted 29 April 2021
Published 27 May 2021

Downloaded from https://journals.asm.org/journal/mra on 18 March 2025 by 131.111.184.102.

TABLE 1 Accession numbers and characteristics of kefir fungal isolates

Species	Strain	SRA accession no.	GenBank accession no.	No. of reads	Coverage (×)	Genome size (bp)	GC content (%)	No. of contigs	Contig N_{50} (bp)	No. of genes	Single-copy BUSCOs (%)
Candida californica	SB-48	SRX9449769	PUHW00000000	25,448,750	192	12,323,006	28.6	1,206	23,604	5,524	91.1[a]
Candida californica	SB-116	SRX9449771	PUHU00000000	25,749,226	194	12,320,729	28.7	981	28,622	5,490	92.0[a]
Kazachstania exigua	SB-178	SRX9449774	PUHR00000000	27,522,278	189	13,507,013	33.3	773	38,581	5,522	96.9[a]
Kazachstania unispora	SB-162	SRX9449773	PUHS00000000	29,185,902	225	12,020,007	32.3	432	60,809	5,464	96.8[a]
Rhodotorula mucilaginosa	SB-353	SRX9449775	PUHQ00000000	19,300,818	89	20,066,154	60.6	416	112,846	7,169	93.4[b]

[a] The lineage data set saccharomycetes_odb10.
[b] The lineage data set basidiomycota_odb10.

Downloaded from https://journals.asm.org/journal/mra on 18 March 2025 by 131.111.184.102.

quality trim the reads (with the following parameter settings: ILLUMINACLIP:TruSeq2-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36). A separate removal step for other contaminants was not performed. The resulting reads were assembled with ABySS v2.02 (9) and SPAdes v3.9.0 (10). No correction steps were performed for the ABySS assemblies; however, mismatches and short indels were corrected for the SPAdes assemblies (enabled with the --careful flag). The obtained genome assemblies based on different parameters were evaluated based on contiguity and completeness with single-copy orthologs using the QUAST v4.1 (11) and BUSCO v3 (12) tools, respectively. The lineage data sets in the benchmarking universal single-copy ortholog (BUSCO) analysis were saccharomycetes_odb9 (*C. californica*, *K. exigua*, and *K. unispora*) and basidiomycota_odb9 (*R. mucilaginosa*). The best genome assemblies were obtained with ABySS with k-mer length values (parameter k) set to 45 for *K. exigua* and 67 for *R. mucilaginosa*. Regarding the other three strains, SPAdes with default settings and the --careful flag produced the best assemblies. The contigs shorter than 500 bp were discarded.

Assemblies were annotated for repeat regions and soft masked with the RepeatModeler v1.0.11 (13) and RepeatMasker v4.0.7 (14) tools. The protein-encoding sequences (CDSs) and tRNAs were predicted with the funannotate predict function in funannotate v1.5.3 (15). The predicted genes were functionally annotated based on their protein sequences using the funannotate annotate function in funannotate v1.5.3 (15) from the MEROPS v12.0 (16), MIBiG v1.4 (17), Pfam v32.0 (18), dbCAN v7, and eggNOG v4.5.1 (19) databases. Transmembrane and secreted proteins were annotated using Phobius v1.0.1 (20) and SignalP v4.1 (21). Finally, secondary metabolite biosynthetic gene clusters were identified with antiSMASH v4.2.0 (22). Default parameters were used for all software unless otherwise specified.

Table 1 shows that the five newly isolated strains exhibit a genome size range of 12.02 Mb to 20.07 Mb with an average GC content of 28.6% to 60.6%.

Data availability. The raw reads have been deposited at the NCBI Sequence Read Archive (SRA), and the whole-genome shotgun projects have been deposited at DDBJ/ENA/GenBank. While all these data are available under BioProject number PRJNA435582, the individual SRA and GenBank accession numbers described in this report are included in Table 1. The GenBank versions described in this paper are the first versions (01).

ACKNOWLEDGMENTS

The DNA sequencing libraries were created and sequenced at the EMBL Genomics Core Facility.

This work was sponsored by the German Ministry of Education and Research (BMBF) (grant number 031A601B) as a part of the ERASysAPP project SysMilk.

REFERENCES

1. Lopitz-Otsoa F, Rementeria A, Elguezabal N, Garaizar J. 2006. Kefir: a symbiotic yeasts-bacteria community with alleged healthy capabilities. Rev Iberoam Micol 23:67–74. https://doi.org/10.1016/S1130-1406(06)70016-X.
2. Simova E, Beshkova D, Angelov A, Hristozova T, Frengova G, Spasov Z. 2002. Lactic acid bacteria and yeasts in kefir grains and kefir made from them. J Ind Microbiol Biotechnol 28:1–6. https://doi.org/10.1038/sj/jim/7000186.
3. Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, Glöckner FO. 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res 41:e1. https://doi.org/10.1093/nar/gks808.
4. Nilsson RH, Larsson K-H, Taylor AFS, Bengtsson-Palme J, Jeppesen TS, Schigel D, Kennedy P, Picard K, Glöckner FO, Tedersoo L, Saar I, Kõljalg U, Abarenkov K. 2019. The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. Nucleic Acids Res 47:D259–D264. https://doi.org/10.1093/nar/gky1022.
5. Kowalczyk M, Kolakowski P, Radziwill-Bienkowska JM, Szmytkowska A, Bardowski J. 2012. Cascade cell lyses and DNA extraction for identification of genes and microorganisms in kefir grains. J Dairy Res 79:26–32. https://doi.org/10.1017/S0022029911000677.
6. Blasche S, Kim Y, Mars RAT, Machado D, Maansson M, Kafkia E, Milanese A, Zeller G, Teusink B, Nielsen J, Benes V, Neves R, Sauer U, Patil KR. 2021. Metabolic cooperation and spatiotemporal niche partitioning in a kefir microbial community. Nat Microbiol 6:196–208. https://doi.org/10.1038/s41564-020-00816-5.
7. Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
8. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170.
9. Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L, Warren RL, Birol I. 2017. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome Res 27:768–777. https://doi.org/10.1101/gr.214346.116.
10. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.
11. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality

Downloaded from https://journals.asm.org/journal/mra on 18 March 2025 by 131.111.184.102.

assessment tool for genome assemblies. Bioinformatics 29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086.
12. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212. https://doi.org/10.1093/bioinformatics/btv351.
13. Smit A, Hubley R. 2015. RepeatModeler Open-1.0. https://www.repeatmasker.org/RepeatModeler/
14. Smit A, Hubley R, Green P. 2015. RepeatMasker Open-4.0. https://www.repeatmasker.org/RepeatMasker/
15. Palmer J, Stajich J. 2019. nextgenusfs/funannotate: funannotate v1.5.3. https://doi.org/10.5281/ZENODO.2604804.
16. Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn RD. 2018. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. Nucleic Acids Res 46:D624–D632. https://doi.org/10.1093/nar/gkx1134.
17. Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, de Bruijn I, Chooi YH, Claesen J, Coates RC, Cruz-Morales P, Duddela S, Düsterhus S, Edwards DJ, Fewer DP, Garg N, Geiger C, Gomez-Escribano JP, Greule A, Hadjithomas M, Haines AS, Helfrich EJN, Hillwig ML, Ishida K, Jones AC, Jones CS, Jungmann K, Kegler C, Kim HU, Kötter P, Krug D, Masschelein J, Melnik AV, Mantovani SM, Monroe EA, Moore M, Moss N, Nützmann H-W, Pan G, Pati A, Petras D, Reen FJ, Rosconi F, Rui Z, Tian Z, Tobias NJ, Tsunematsu Y, Wiemann P, Wyckoff E, Yan X, et al. 2015. Minimum Information about a Biosynthetic Gene cluster. Nat Chem Biol 11:625–631. https://doi.org/10.1038/nchembio.1890.
18. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M. 2014. Pfam: the protein families database. Nucleic Acids Res 42:D222–D230. https://doi.org/10.1093/nar/gkt1223.
19. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, Bork P. 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol Biol Evol 34:2115–2122. https://doi.org/10.1093/molbev/msx148.
20. Kall L, Krogh A, Sonnhammer ELL. 2005. An HMM posterior decoder for sequence feature prediction that includes homology information. Bioinformatics 21:i251–i257. https://doi.org/10.1093/bioinformatics/bti1014.
21. Nielsen H. 2017. Predicting secretory proteins with SignalP. Methods Mol Biol 1611:59–73. https://doi.org/10.1007/978-1-4939-7015-5_6.
22. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, Weber T, Takano E, Breitling R. 2011. antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res 39:W339–W346. https://doi.org/10.1093/nar/gkr466.

Downloaded from https://journals.asm.org/journal/mra on 18 March 2025 by 131.111.184.102.